Genre Specific Classification for Information Search and Multimodal Semantic Indexing for Data Retrieval

S. Gomathy*, K.P. Deepa**, T. Revathi*** & L. Maria Michael Visuwasam****

*Student, Department of Computer Science & Engineering, Velammal Institute of Technology, Panchetti, Tamilnadu, INDIA.
E-Mail: gomathysiva@ymail.com
**Student, Department of Computer Science & Engineering, Velammal Institute of Technology, Panchetti, Tamilnadu, INDIA.
E-Mail: prodigydeepy@yahoo.co.in
***Student, Department of Computer Science & Engineering, Velammal Institute of Technology, Panchetti, Tamilnadu, INDIA.
E-Mail: revathithangamuthu@ymail.com
****Assistant Professor, Department of Computer Science & Engineering, Velammal Institute of Technology, Panchetti, Tamilnadu, INDIA.
E-Mail: micael_vm@yahoo.co.in

***Abstract*—**Searching has become an integral part in today's Internet bound era. Now-a-days, searching a video content from a large scale video set is difficult because the volume of video increases rapidly that too with the lack of proper tools to handle. Large video collections such as YouTube contains many different genres that searches through Tree- Based Concept and retrieves video result through filename that are subjective and noisy and in many cases not reflecting the accurate content. The overall aim of the paper is to provide ease to the user by giving a refined video search by accepting both text and image inputs. Our system describes architecture for a new video searching mechanism that not only accepts text based inputs but also accepts image based input from the user to retrieve the video results. Here we propose, two step frame work comprising of two levels, genre level and semantic concept level, with a filtering mechanism to generate the accurate video result.

***Keywords*—**Frame Matching, Genre Specific Classification, Image based Search, Multimodality Web Categorization, Multiple Frame Detection, Semantic Indexing and Video Retrieval

***Abbreviations*—**Content-Based Video Retrieval (CBVR), Red-Green-Blue Scale Invariant Feature Transform (RGB-SIFT), Support Vector Machine (SVM), Text Retrieval Conference Video Retrieval Evaluation (TRECVID)

I. INTRODUCTION

THE Web search has become an indispensable part of our lives as they allow us to access information from anywhere at any time. Online searching mechanism enables quick and inexpensive research investment opportunity from virtually anywhere. The web based search is always time consuming and convenient for any user. Unfortunately in terms of large set videos, the searching of contents has been at times overshadowed by retrieving unwanted results which is not needed by the user. If we were able to reliably provide descriptive labels [Worring et al., 2011] from a large-scale video set, we could improve the ease and speed with which we access and manage videos significantly.

Our work has effectively solved classical learning problem in the existing system leads to results unwanted results. We do so by using a two step framework that uses multimodality web categorization algorithm. By considering the different semantic levels [Jun Wu & Marcel Worring, 2012] between different genres, searching of video contents is accomplished for both text and image based inputs from the user.

The main goal of our paper is to use Multimodality web Categorization Algorithm which considers detecting genre specific semantic concepts for a given video by using two stages. In the first stage, the video genre models are trained using efficient global features and then the genre specific concepts models are trained using complex, local and object level features.

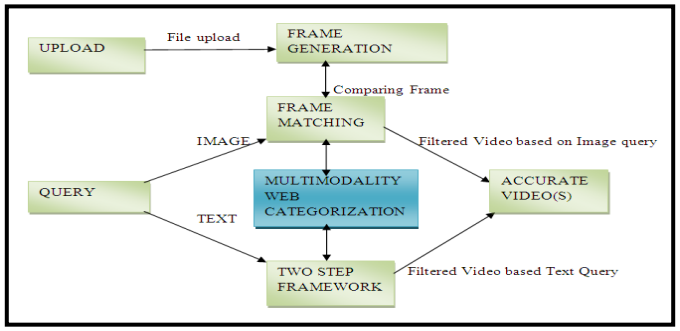

Figure 1 – Overall Structure of the Concepts

A framework for the classification of feature into genres, based on computable visual cues proposed by Rasheed et al., (2005) classifies the movies into four broad categories and four computable video features(average shot length, color variance, motion content and lighting key) that are combined in a framework to provide a mapping to these four high-level semantic classes. Here mean shift classification is used to discover the structure between the computed features. In our paper, we have proposed the same approach of category division by using the two step framework where the video is computed into frame objects and the features such as pixel calculation which is done with array lists and hash mapping involves minimal human intervention.

The utility of a fixed lexicon of visual semantic concepts for automatic multimedia retrieval and re-ranking purposes explores [Natsev et al., 2007] the several new approaches for query expansion, in which textual keywords, visual examples or initial retrieval results are analyzed to identify the most relevant visual concepts for the given query. These concepts are then used to generate additional query results. The results are evaluated using a large video corpus and 39 concept detectors from the TRECVID-2006 video retrieval benchmark. In our paper, we develop both lexical and statistical approaches for text query expansion as well as content-based approaches for visual query expansion by using the specified automatic multimedia retrieval and observe consistent improvement relative to a state of -the-art multimodal retrieval baseline. The concept of cataloging a large component of the web contents that consists of visual information such as images, graphics, animations and videos which requires a highly efficient automated system that regularly traverses the web, detects visual information, processes it and indexes it in such a way as to allow efficient and effective search retrieval [Naphade et al., 2000]. But since the visual information on the web proves to be highly volatile, our paper not only involves in indexing the information as specified but also object level filtering is done.

Content-Based Video Retrieval (CBVR) systems that are amenable to support automatic low-level feature extraction is discussed by Fan et al., (2009). A novel multimodaling boosting algorithm is proposed by incorporating feature hierarchy and boosting to reduce the training cost significantly [Worring et al., 2011]. To bridge the semantic gap between the available video concepts and the user's real needs, a novel hyperbolic visualization framework is seamlessly incorporated to enable intuitive query specification. Our paper uses the above specified multimodality approach in two stages where effective filtering is each stage which removes complex correlations and retrieve semantic results.

Image category recognition that is important to access visual information on the level of objects and scene types is dealt [Van de Sande et al., 2010]. To increase illumination invariance and discriminative power, color descriptors have been proposed. The invariance properties and the distinctiveness of color descriptors have been explored using taxonomy with respective to photometric transformation. From the experimental results it is derived that invariance to light intensity changes affects category recognition. Our paper uses the above concept by involving in pixel calculation characterized by width, height, light intensity which increases the effect of visual information.

II. RELATED WORKS

2.1. *Multiple Frame Detection*

2.1.1. *Spatio Temporal Selection*

Every shot changes due to object motion and camera motion. This fact requires us to divide the video content into multiple frames where each frame has minute difference than the other. The Spatio-temporal selection [Snoek et al., 2010] makes it possible to recognize the concepts by taking more frames into account that are visible during the shot, but not necessarily in a single frame. Hence our approach has a strong dependency on the spatio-temporal concept that analyses multiple frames to obtain higher performance in the visual appearance of a semantic concept over the previously used key frame methods.

2.1.2. *Probabilistic Frame Division*

To be precise, we in our approach sample n number of frames depending on the length and duration of the video. We have used the concept of assigning different probability based on the object motion and the duration of visual content. The video genres in the training set is assigned with the character set V where this set consists of n number of frames associated with each video. V= $\{k_{i1},k_{i2},k_{i3}.....\}$ where k denotes each video in the training set and i denotes the number of frames that each video consists of depending on their duration 'i' is calculated by

$$p(k_{i1}|v) = \frac{p(j|k_1)p(j|k_2)p(j|k_3)}{\sum M(V_{ij})}$$

Where j denotes the duration of each video and M denotes the total common factor based on the visual content of the training set.

2.2. *Automatic Web Video Categorization*

The related videos frequently have relevant contents or similar category labels with the given video. At the same time, users share videos based on their personal interests, and

therefore the uploaded videos by the same user usually have similar type. By integrating the contextual and social information to effectively categorize web videos [Wu et al., 2009] semantic meaning, video relevance, and user interest, respectively, are fused together to obtain the probability score and predefined category based on the SVM classifier. Borth, J. Hees & Koch (2009) in their approaches, proposed the technique of automatic web categorizer which inserts videos into a Genre- Hierarchy using statistical classification based on tags, Titles and visual features. This approach puts a strong weight on the classification from the meta information. Visual categorization uses the well known "bag of visual words" model where clips are represented by key frames that supports rich genre hierarchy. This "bag of visual words" model also called bag of-features representation [Marszałek et al., 2009] views videos as spatio-temporal volumes. Finally for scalable training, efficient on-line learning strategies are investigated.

III. METHODS

3.1. Video Concept Detection

3.1.1. Concept Extraction from a Visual Content

The detected number of frames for each video set is subjected to object level filtering and treated globally as efficient complex objects. Each frame is stored as a single image thus lessening the complexity of video retrieval and emphasis the search both locally and globally with minimal human intervention. Based on the previous TRECVID 2010 experiments and media mill semantic video search engines [Snoek et al., 2010] it has been approved that bag- of -words approach projected by Schmidt has made the system more efficient to provide a standardized input-output model. To allow for semantic integration, the media mill TRECVID 2010 consists of supervised machine learning process that is composed of two phases-training and testing. The optimal configuration of features is learned from the training data in the first stage and probability p(j|ki) is assigned to each input feature vector for each semantic concept by using the classifier in the second stage. In our approach we expand the training set into a two step framework that broadens the categorization into many semantic and structural levels. The training set acknowledges for unwanted data level filtration based on the genres [Xu & Li, 2003]. Here we set to use the MPEG-7 compliant low-level content descriptors on color and textures as part of the visual content description feature, which include Scalable Color, Layout and Texture.

3.2. Detecting Sift Features

Snoek et al., (2010) analyzed the classes within the diagonal model of illumination change and specifically data sets considered within TRECVID deals with SIFT, Opponent SIFT and RGB-SIFT to observe the changes in the light intensity for detecting the differences in each single frame of visual content. We make use of RGB-SIFT in our approach that hypothesis the normalizations due to invariance of the color shift and recognize the light intensity of each pixel in the multiple frames. The RGB-SIFT calculates the length and width of each pixel based on the clustering invariance algorithm. This shift advances the color shift features which make use of the code book and the K-Means clustering algorithm.

Specifically, with the use of Spatial-Pyramid Matching technique [Li et al., 2010] to represent the low-level visual local features, SIFT and MoSIFT is used to describe the location information of these feature points. Another detector training part besides the traditional SVM is the Sequential Boosting SVM classifier which is proposed to deal with the large-scale unbalanced data classification problem. The cross-validation process [Yuan et al., 2006] divides the training part into two: one is used to train a SVM classifier and the other is used to test the obtained classifier. For multimedia event detection and semantic indexing of concepts the main tasks involve extracting the features, computing SIFT in them.

3.3. Multimodality Web Categorization

Our framework uses a semi –automatic fashion by suggesting category for the User's clip. These categories are subdivided into a broaden approach which involves sub-categorization based on the selected category. The Multimodality Web Categorization approach accounts for selecting the proper sub–category based on the desired category. Here quick object level filtering is responsible for indexing the genres based on the semantic and visual concepts. This multimodal approach [Snoek & Worring, 2005] in which either the most appropriate modality is selected or the different modalities used in collaborative fashion is the backbone of effective indexing.

3.4. Semantic Indexing

Many kinds of content-based video navigation is supported and utilized by thousands and hundreds of semantic features [Smeaton et al., 2009]. Our approach relies on matching one video frames against other using low level character such as texture, color or determining and analyzing the object appearing within the video. We have used a hypothetical indexing technique that analyses a fixed vocabulary set. This set consists of global and local synonyms that are the words with relative meaning, to analyze the visual content so called as high level features also referred to as semantic concept. Low level features for an object assigns for the object's description at a higher semantic level that is better attuned to matching information needs. This technique has been more efficient in a compositional fashion to reduce the human intervention, and broaden the search selection to the relevant shots.

The semantic meaning of text (title and tags), and video relevance from related videos, are induced to robustly determine the video category. The common vocabulary set consists of all related words that would retrieve the appropriate video for the relevant text query. This indexing

stimulates and evaluates the list of individual words that contributes to the same or approximate meaning.

The related videos frequently have relevant contents or similar category labels with the given video. At the same time, users share videos based on their personal interests and therefore the uploaded videos by the same user usually have similar type [Song et al., 2009]. By integrating the contextual and social information to effectively categorize web videos, semantic meaning, video relevance, and user interest respectively are fused together to obtain the probability score and predefined category based on the SVM classifier.

IV. RESULTS

Looking at the experimental results of our paper, we first consider the semantic indexing concept. During the search mechanism by a user, the videos are retrieved based on the user entered query. That is, the user entered query is compared with the file name and the description of a particular video. For example : if a video related to babies is being uploaded by an user with the filename as "Baby" and description "this is a video of baby playing in water "and at anytime there may be some other user who wants to view this video. When this user enters the query in the search bar as "infant videos ", now this is where evaluating the efficiency is revealed. With the help of concept in which the alternate word for infants, which are "child", "kid", "baby", "toddler", has already been stored in the database using the SEMANTIC INDEXING CONCEPT, based on which all the videos with the filename or description which has the word" child", "kid", "baby", "toddler", "infants" in it will be exactly retrieved.

Nextly uploading a video is achieved by mentioning the file path, category and the sub-category that the file belongs to and also a description about the video. Only if the category and sub-category matches, an authenticated user will be able to upload the video. For example: if the user mentions the category as entertainment and then chooses the sub-category as Education or any such sub-categories that are not related to entertainment, then the video will not uploaded and an error message is prone to be thrown. Example 2: similarly each category like Entertainment, Education, Sports mentioned by the person who is uploading should always be followed by the correct and valid sub-category like Entertainment-movies, Entertainment-dance, Entertainment-songs etc for entertainment category and holds the same for education as education-medical, education-engineer, education-arts and for sports as Sports-hockey, sports-tennis, sports-cricket, etc.

In Frame Generation once the videos are uploaded, the corresponding frames for those videos are created and are stored in separate folders with the name same as the video file name. Example: If the video is being uploaded with the name "National Flag" then at the time of uploading the video, a corresponding folder with the same name as the video name, that is "National Flag" is created, which consists of the generated frames of the video that is being uploaded. The number of frames generated solely depends on the duration of the video that has been uploaded.

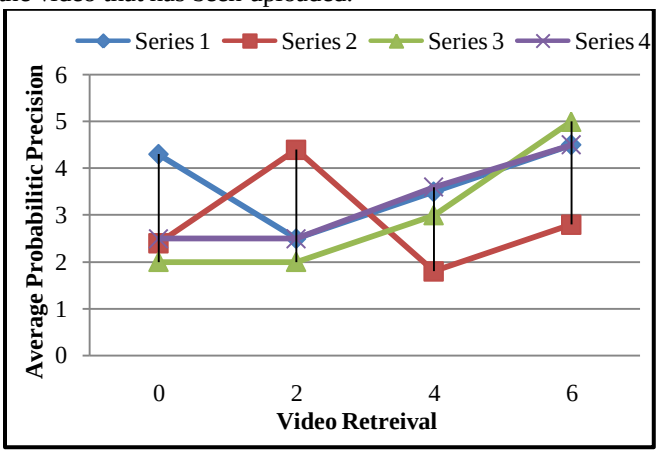

Figure 2 – Accurate Frame Retrieval

An additional feature that has been included is the frame matching, which is implemented when the same video is being uploaded twice. When a same video is being uploaded twice, at the time of upload, the system checks for all the frames generated for each video and eliminate the uploading of same videos twice, in case any frame of the uploading video matches with the frame of the already uploaded video. The frame matching and elimination of uploading the same video twice is purely content based.

An important factor of web based search is that when the users enter the query in the query search bar, even without mentioning the category and sub-category the exact results will be retrieved based on the description and the semantic indexing concepts that has been inculcated during the later stages of the uploading a file. Example: if a user is searching for a sports video, user can just type the text based query in the search bar and mentioning the category and sub-category is not mandatory, but still the user will be directed with the exact video result.

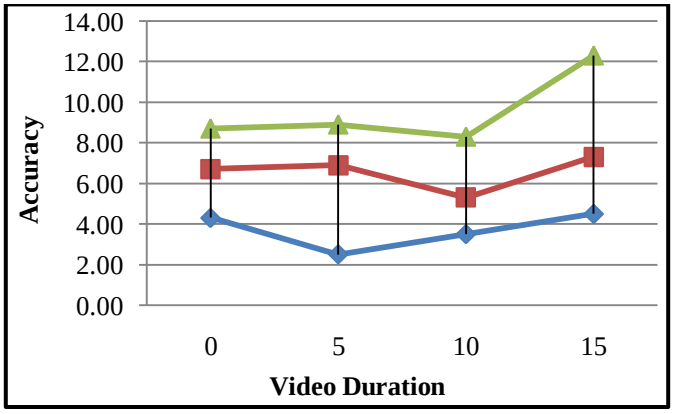

Figure 3 – Frame matching Accuracy

In Image input search, the user can give an image input in the search bar and the frames of all the videos is being traversed and checked to fetch the video that contains the exact frame (image) that the user has given in the search bar. Example: suppose a user gives a input image in the search bar as an Indian flag, but there are many folders which has already been generated for each video at the time of its

upload, the mechanism adapted allows the search to be refined and so it identifies the accurate folder that contains exactly the same frame as that of the image given in the search bar and as an end result, that corresponding video is being retrieved and displayed to the user. Even when a frame in a folder of a particular video is being duplicated and mixed or pasted with the frames of the other videos, still the exact video result will be retrieved. Example if there are four videos uploaded, one a video of Indian flag, then a video of a dance, then video of a baby and then a video of a comics. When the generated frames of all these videos is being mixed and kept in any one folder, say the dance folder. Now suppose a user gives the input image in the search bar as a frame generated from flag video, then the output displayed to the user will be the Flag video only and not the other videos. Since the search mechanism searches for the folder that contains the frames similar to the "flag" which is given as an image input in the search bar, the corresponding video is retrieved proving the image-based search to be highly content based. The pixel and brightness information of the image given as input in the search bar is same as that of pixel, brightness of the video that is retrieved as the exact result.

V. CONCLUSION AND FUTURE WORK

Video genre classification is applied first to filter out most of the irrelevant material resulting in the relatively small subset. Then the genre specific concept or topic models are applied. Our proposed work provides ease to the user to retrieve accurate video results by accepting both text and image based inputs. The implemented system effectively minimizes the retrieval of unrelated video sets to the users, who search for a significant content. It adapts a two-step framework, a genre specific classification for information search and multimodal semantic indexing, for data retrieval. This framework is achieved for both the text-based user input, as well as image based user input for searching and retrieving the exact video content in less time efficiently, thereby avoiding the retrieval of number of unrelated videos, that the significant number of users searches. This is an advance to the existing approach that uses automatic web categorization algorithm, which attempts to retrieve even the unrelated information or video, for a user search query, making the search handling mechanism more complex. Our searching could be implemented to handle any video search and this implemented system works well to block all the possibilities of occurrence of errors, adapting a high robust filtering operation with a minimal complexity. Thus revealing a pleasing appearance to all the genuine users who relishes the ease of using our concepts. Our implemented system can be extended to explore the visual content for even a text based user input during the query search. We might extend this approach based on Tag-localization to localize the tags associated with each video shots and estimating the relevance scores of these shots with respect to the query. Then to create a set of key shots which should be identified by performing near-duplicate key frame detection. We also aim at using many special tags like geo-tags and event-related tags that can be explored to further improve the performance of our approach in such a way that the videos searched by any user can be better tracked. We leave them to our future work.

REFERENCES

[1] M.R. Naphade, I. Kozintsev & T. Huang (2000), "Probabilistic Semantic Video Indexing", *Proceedings of Neural Information Processing Systems Conference*, Pp. 967–973.

[2] L.-Q. Xu & Y. Li (2003), "Video Classification using Spatial-Temporal Features and PCA", *Proceedings of International Conference on Multimedia and Expo*, Pp. 485–488.

[3] G.M. Snoek & M. Worring (2005), "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools Applications*, Vol. 25, No. 1, Pp. 5–35.

[4] Z. Rasheed, Y. Sheikh & M. Shah (2005), "On the use of Computable Features for Film Classification", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 1, Pp. 52–64.

[5] X. Yuan, W. Lai, T. Mei, X. Sheng Hua (2006), "Automatic Video Genre Categorization using Hierarchical SVM", *Proceedings of International Conference on Image Processing (ICIP)*, Pp. 2905–2908.

[6] P. Natsev, A. Haubold & J. Tešić (2007), "Semantic Concept-based Query Expansion and Re-Ranking for Multimedia Retrieval", *ACM Transactions on Multimedia*, Pp. 991–1000.

[7] X. Wu, W.-L. Zhao & C.-W. Ngo (2009), "Towards Google Challenge: Combining Contextual and Social Information for Web Video Categorization", *Proceedings ACM Multimedia*, Pp. 1109–1110.

[8] M. Marszałek, I. Laptev & C. Schmid (2009), "Actions in Context", *Proceedings Conference on Computer Vision and Pattern Recognition*, Pp. 2929–2936.

[9] Y. Song, Y.-D. Zhang & X. Zhang (2009), "Google Challenge: Incremental-Learning for Web Video Categorization on Robust Semantic Feature Space", *Proceedings ACM Multimedia*, Pp. 1113–1114.

[10] Borth, J. Hees & M. Koch (2009), "Tubefiler: An Automatic Web Video Categorizer", *Proceedings ACM Multimedia*, Pp. 1111–1112.

[11] F. Smeaton, P. Over & W. Kraaij (2009), "High-Level Feature Detection from Video in Trecvid: A 5-Year Retrospective of Achievements", Editor: A. Divakaran, *Multimedia Content Analysis, Theory and Applications*, Berlin, Germany: Springer-Verlag.

[12] J. Fan, H. Luo & Y. Gao (2009), "Incorporating Concept Ontology for Hierarchical Video Classification, Annotation, and Visualization", *IEEE Transactions on Multimedia*, Vol. 9, No. 5, Pp. 939–957.

[13] K.E. Van de Sande, T. Gevers & C.G. Snoek (2010), "Evaluating Color Descriptors for Object and Scene Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, Pp. 1582–1596.

[14] H. Li, L. Bao, Z. Gao & A. Overwijk (2010), "Informedia At TRECVID 2010", *Proceedings TRECVID Workshop*.

[15] G. Snoek, K.E. Van de Sande, O. de Rooij (2010), "Themediamill TRECVID 2010 Semantic Video Search Engine", *Proceedings TRECVID Workshop*.

[16] M. Worring, Ulges, & T. Breuel (2011), "Learning Visual Contexts for Image Annotation from Flickr Groups", *IEEE Transactions on Multimedia*, Vol. 13, No. 2, Pp. 330–341.

[17] Jun Wu & Marcel Worring (2012),"Efficient Genre-Specific Semantic Video Indexing", *IEEE Transactions on Multimedia*, Vol. 14, No. 12, Pp. 33–38.

L. Maria Michael Visuwasam, born on 5th September, 1981 near Kovilpatti. He received the B.Tech degree in Information Technology from Anna University, Chennai with Distinction in 2005. He has received M.E degree in Computer Science and Engineering specialization with Knowledge Engineering from College of Engineering, Anna University, Chennai, India in 2008. He received MBA degree in Education Management from Alagappa University, Karaikudi, India in 2010 and registered Ph. D degree in Anna University, Chennai in 2010. He is doing research in the area of Music Emotion Recognition. He has been with the Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai as Assistant Professor. He has published two international journals, six international conferences, ten national conference papers and got best paper award from NCRTIS-2K11. His research includes emotion recognition, Ad-hoc Networks and web security.

S. Gomathy, born on 21th September 1992 in Chennai. She is currently pursuing her B.E computer science and engineering in Velammal Institute of Technology, Chennai. She has successfully completed a certificate course in Oracle PL/SQL programming conducted in Ramanujam Computing Centre, Anna University during Jan2012-March2012.She had also participated in IBM Great Mind Challenge Competition and presented a web application project on Picture Geotag Survey in June 2012.

K.P Deepa, born on 20th July 1991 in Chennai. She is currently pursuing her B.E computer science and engineering in Velammal Institute of Technology, Chennai. She has successfully completed a certificate course in Oracle PL/SQL programming in Anna University during Jan2012-March2012.She had also participated in IBM Great Mind Challenge Competition and presented a web application project on Picture Geotag Survey in June 2012.

T. Revathi, born on 6th July 1991 in Chennai. She is currently pursuing her B.E computer science and engineering in Velammal Institute of Technology, Chennai. She has successfully completed her Microsoft Certification and qualified as Microsoft Certified Technology Specialist in .NET Framework 3.5; ASP.NET Applications in 2011.She had also participated in IBM Great Mind Challenge Competition and presented a web application project on Picture Geotag Survey in June 2012.